Phylogenetic shadowing of primate sequences to find functional regions of the human genome

Dario Boffelli[1,2], Jon McAuliffe[3], Dmitriy Ovcharenko[2], Keith D. Lewis[2], Ivan Ovcharenko[2], Lior Pachter[4] and Edward M. Rubin[1,2,*]

[1] DOE Joint Genome Institute

Walnut Creek, CA 94598

[2] Department of Genome Sciences,

Lawrence Berkeley National Laboratory

Berkeley, CA 94720

[3] Department of Statistics

University of California, Berkeley

Berkeley, CA 94720

[4] Department of Mathematics

University of California, Berkeley

Berkeley, CA 94720

[*]to whom correspondence should be addressed at: emrubin@lbl.gov

Abstract

Non-human primates represent the most relevant model organisms to understand the biology of *Homo sapiens*. The recent divergence and associated overall sequence conservation between individual members of this order have nonetheless largely precluded the use of primates in comparative sequence studies. We used sequence comparisons of an extensive set of Old- and New-World monkeys and hominoids to identify functional regions in the human genome. Analysis of this data enabled the discovery of primate-specific gene regulatory elements and the demarcation of the exons of multiple genes at a greater resolution than attainable by human-mouse comparisons. Much of the information content of the comprehensive primate sequence comparisons could be captured with a small subset of phylogenetically close primates. These results demonstrate the utility of intra-primate sequence comparisons to discover primate-specific as well as common mammalian functional elements in the human genome, unattainable through the evaluation of more evolutionarily-distant species.

Genomic sequence comparisons between distant species have been extensively used to identify genes and determine their intron-exon boundaries, as well as to identify regulatory elements present in the large non-coding fraction of the genome (*1-3*). This strategy has been successful in human/mouse comparisons, as the approximately 75 million years (MY) of separation from their last common ancestor have provided sufficient time for a large fraction of nucleotides to have been exposed to considerable mutation and selection pressure. While such comparisons readily identify regions of the human genome performing general biological functions shared with evolutionarily distant mammals, they will invariably miss recent changes in DNA sequence that account for uniquely-primate biological traits.

As a consequence of their short evolutionary separation (chimp 6 MY, Old World monkeys 25 MY, New World monkeys 40 MY) (*4*), there is a paucity of sequence variation between human and each of its nearest primate relatives. This makes it difficult to distinguish functional from passive conservation on the basis of pairwise comparisons, thus limiting the usefulness of such comparisons. However, the additive collective divergence of higher primates as a group (Fig. S1) is comparable to that of humans and mice. This suggests that deep sequence comparisons of numerous primate species should be sufficient to identify significant regions of conservation that encode functional elements. Phylogenetic footprinting (5,6) is a method which has been used to identify highly conserved putative *cis*-acting regulatory elements exploiting alignments across numerous evolutionarily-distant species. We developed a variant of phylogenetic footprinting, which we termed phylogenetic shadowing. In contrast to footprinting,

phylogenetic shadowing examines sequences of closely-related species and importantly takes into account the phylogenetic relationship of the set of species analyzed. This approach enabled the localization of regions of collective variation and complementary regions of conservation, facilitating the identification of coding as well as non-coding functional regions.

We first examined the ability of this strategy to identify functional regions with precise locations within the human genome, such as intron-exon boundaries. The lack of clone-based libraries for multiple primate species limited us to sequencing orthologous regions from a large set of primates using genomic DNA as template (*5*). The sole criterion used in the selection of the four different regions we studied was that each contain at least one annotated exon. The sequences were generated (*6*) for a set of 13-17 primate species that included those evolutionarily closest to human, such as Old- and New-World monkeys and hominoids, but not distant primates such as prosimians. The resulting sequences were analyzed to determine the likelihood ratio under a fast- versus a slow-mutation regime for each aligned nucleotide site across all four regions analyzed *(6)* (supporting online text). This represents the relative likelihood that any given nucleotide site was subjected to a faster or slower rate of accumulation of variation and is related to functional constraints imposed on each site. The corresponding likelihood ratio curves were used to describe the variation profile of the four genomic intervals analyzed.

In all regions examined, the exon-containing sequences displayed the least amount of cross-species variation, in agreement with the constraint imposed by their functional role

(Fig. 1, Panels A-D). A limited number of short regions of minimal variation similar to the exon-containing sequences appeared in the likelihood plots. These regions, however, were all less than 50 bp long, making them unlikely to be candidate exons for alternative splicing (the vast majority of internal exons predominantly range between 50 and 200 bp in length (*7*)). These regions may represent previously unidentified regulatory elements. In agreement with observations from other parts of the genome (*8*), the four sequenced regions have evolved at different rates, as indicated by their differing absolute likelihoods. Comparison of the human-mouse versus the multiple-primate visualization plots (Fig. 1, Panels a-d) yields similar results with exons showing the highest level of conservation in all regions studied. The primate sequence comparisons illustrate the effectiveness of the phylogenetic shadowing method in yielding a precise identification of the exon boundaries. In the case of the cholesteryl ester transfer protein gene (Fig. 1, Panels B and b), the human-mouse comparison is unavailable, since this gene has been inactivated in the mouse. This inactivation represents an extreme example of the frequently encountered situation where the mouse genome, due to the lack of meaningful alignments between human and mouse sequences, fails to localize functional elements in the human genome (*9*).

We next analyzed the sequence data for the four regions studied to assess the contribution of additional species to the discriminative power of phylogenetic shadowing and identify the most informative minimum subset of species *(6)*. The sequences from the most informative subset of only four to seven species, depending on the genomic locus, were

calculated to be sufficient to capture approximately 75% of the total available discriminative power of this approach (Table 1). We were able to unequivocally identify the position exon 3 of liver-X-receptor-α from the five most informative species (Homo, Saguinis, Colobus, Callicebus, Allenopithecus) , as indicated by the likelihood curve (Fig. S2). Comparison with Fig. 1C shows that additional species only marginally improve this plot. As would be predicted, the species included in the set maximizing the discriminative power of phylogenetic shadowing include representatives of the different subfamilies that compose the primate phylogenetic tree and therefore constitute the least-related species in the set studied (Fig. S1).

To investigate the ability of phylogenetic shadowing to discover primate-specific gene regulatory elements, we studied apolipoprotein(a) (apo(a)), a recently evolved primate gene whose orthologous distribution is limited to Old World monkeys and hominoids (*10*). Defining the regulatory sequences determining apo(a) expression levels is of considerable biomedical relevance, since high plasma levels of this protein serve as an important cardiovascular disease risk predictor (*11, 12*). The sparse distribution of this gene among mammals precludes classical comparative genomic approaches. Sequence comparison of the apo(a) locus in humans, chimps and baboons revealed a region of extreme conservation of approximately 1.6 kb adjacent to the transcription start site, surrounded by approximately 8 kb of reduced conservation on either side owing to the different patterns of repetitive-element insertion in the three species. We sequenced and analyzed this 1.6 kb region in 18 Old World monkeys and hominoids. Phylogenetic

shadowing revealed regions of varying conservation similar to that observed for the intron-exon regions described previously (Fig. 2A). Besides the region containing apo(a)'s first exon (region E) (*13*), the remaining regions with the lowest variation include the apo(a) promoter's TATA box (region C9) and a critical HNF-1α transcription factor binding site (region C10), which have both been functionally described (*14*). Eight additional regions, varying in size between 40 and 70 bp, show levels of conservation comparable to these three functional regions, suggesting they are also biologically important.

Since these elements were located immediately upstream of the apo(a) promoter, we tested these sequences for their ability to be recognized by DNA-binding proteins. The highly-conserved regions were predicted to interact with such proteins more efficiently than regions characterized by a low degree of conservation. We performed electrophoretic mobility shift assays on ten oligonucleotides spanning the most-conserved regions and, as putative negative controls, seven similarly-sized oligonucleotides representing regions with the least conservation (Regions N1-7, Fig. 2A). In this assay, nuclear extracts from a liver cell line were mixed with radiolabeled oligonucleotides and separated by electrophoresis *(6)*. All oligonucleotides from the conserved regions interacted with one or more DNA-binding proteins, as reflected by the slower electrophoretic mobility of the protein-bound oligonucleotides relative to the mobility of the pure oligonucleotides (lower bands in Fig. 2B). Conversely, oligonucleotides from non-conserved regions showed only weak or no protein binding. Overall,

oligonucleotides from conserved regions interacted greater than six-fold more strongly with hepatic nuclear proteins than regions with lower conservation (Regions C1-C10.2 vs. N1-N7, Fig. 2C).

Consistent with the prediction that conserved regions interact preferentially with DNA-binding proteins is the prediction that these regions are involved in transcriptional regulation activity. To explore this, we developed an assay for *in vitro* enhancer activity based on transient transfection of a human liver cell line. The ten most-conserved regions, as well as the seven least-conserved regions, were individually deleted from the whole 1.6 kb region *(6)*. These constructs were compared to the intact 1.6 kb region construct for their ability to drive the expression of a luciferase reporter gene. Three independent experimental determinations reproducibly indicated that conserved regions had a significantly larger functional impact than non-conserved regions in these reporter gene expression assays *(6)*. The deletions of the conserved sequence elements reduced the expression of the full reporter by 25% to 55% (*15*), with the exception of the construct carrying the deletion of region 6 (Fig. 3). In contrast, deletions of all but one of the seven non-conserved regions had a minimal impact on the expression of the full reporter construct, reducing its levels by a median value of only 4%.

The analysis of closely-related primates facilitates the discovery of functional elements specific to primates as well as elements shared with evolutionarily-distant mammals. The high absolute degree of similarity minimizes ambiguity in the computation of the

multiple alignment, which in turn greatly facilitates the subsequent construction of the phylogenetic tree (*16*). Furthermore, the generalization of gene-finding algorithms to multiple organism annotation is simplified, which results in more accurate predictions (supporting online text). The facility of sequence alignments and the possibility of comparative assembly for non-human primates using human as the reference sequence have important practical relevance, since they considerably diminish the depth of sequence coverage required for an organism to be informative in annotating the human genome.

Evaluation of our data suggests that sequence from as few as four to six primate species in addition to human is sufficient for the identification of a large fraction of functional elements in the human genome, many of which are likely to be missed by human-mouse comparisons. With the goal of annotating the human genome, these studies indicate that the generation of a limited amount of sequence from a few additional properly-selected primate genomes will provide information previously unavailable from comparisons of humans with more evolutionarily distant species.

References:

1. G. G. Loots *et al.*, *Science* **288**, 136 (2000).
2. L. A. Pennacchio *et al.*, *Science* **294**, 169 (2001).
3. L. A. Pennacchio, E. M. Rubin, *Nat Rev Genet* **2**, 100 (2001).
4. M. Goodman, *Am J Hum Genet* **64**, 31 (1999).
5. Genomic DNA from the following species were analyzed in this study: Pan troglodytes (*), Pan paniscus (*), Gorilla gorilla (*), Pongo pygmaeus (*), Hylobates lar lar (#), Hylobates syndactylus (#), Papio hamadryas, Macaca nemestrina (*), Macaca mulatta (*), Mandrillus leucophaeus (#), Cercopithecus

hamlyni (#), Lophocebus albigena (#), Allenopithecus nigroviridis(#), Miopithecus talapoin (#), Cercopithecus aethiops (#), Erythrocebus patas (*), Colobus guereza kikuyuensis (#), Pygathrix nemaeus (#), Trachypithecus francoisi (#), Presbytis entellus (#), Nasalis larvatus (#), Ateles geoffrey (#), Callicebus moloch (#), Alouatta seniculus (#), Saimiri sciureus (#), Aotus trivirgatus (#), Callithrix (#), Saguinus labiatus (#). DNA for Papio Hamadryas was from the RPCI-41 Baboon BAC Library (www.chori.org). DNA for species marked with * were from the Coriell Cell Repositories (http://locus.umdnj.edu). DNA for species marked with # were kindly provided by the San Diego Zoo/Center for the Reproduction of Endangered Species.

6. Materials and Methods are available as supporting material on Science online
7. E. S. Lander *et al.*, *Nature* **409**, 860 (2001).
8. I. Ebersberger, D. Metzler, C. Schwarz, S. Paabo, *Am J Hum Genet* **70**, 1490 (2002).
9. Mouse Genome Sequencing Consortium, *Nature* **420**, 520 (2002).
10. R. M. Lawn *et al.*, *J Biol Chem* **270**, 24004 (1995).
11. J. Danesh, R. Collins, R. Peto, *Circulation* **102**, 1082 (2000).
12. G. Luc *et al.*, *Atherosclerosis* **163**, 377 (2002).
13. This region was used to learn the mutation rates for "conserved" and "non-conserved" regimes used in this paper.
14. D. P. Wade *et al.*, *J Biol Chem* **272**, 30387 (1997).
15. Deletion of region 9, containing the promoter's TATA box, reduces expression by only 30%, likely attributable to the presence of one of the several cryptic TATA boxes further upstream.
16. C. Notredame, *Bioinformatics* **17**, 373 (2001).

Supporting Online Material
www.sciencemag.org
Materials and Methods
Supporting text
Figs. S1, S2

We would like to thank J.-F. Cheng for support with the sequencing infrastructure, the Zoological Society of San Diego for providing primate DNA samples, Irina Udalova and Len Pennacchio for useful discussions. Michael Jordan contributed useful suggestions concerning the statistical methods we employed. This work was performed under the auspices of the U.S. Department of Energy's Office of Science, Biological and Environmental Research; by the University of California, Lawrence Berkeley National Laboratory under Contract No. DE-AC0376SF00098; supported by the Grant # HL66728, Berkeley-PGA, under the Programs for Genomic Application, funded by National Heart, Lung, and Blood Institute, USA. L.P. was partially supported by a grant from the NIH (R01-HG02362-01).

Table 1. The discriminative power of phylogenetic shadowing at increasing species subset sizes. Each row is an exonic region under study. Each column shows, by region, the percentage of total phylogenetic divergence in complete species set of that region that is captured by the most-divergent subset of the indicated size. The minimum number of species required to capture at least 75% of the full power of the approach is highlighted in bold for each gene.

	Subset Size (number of species)							
Gene	1	2	3	4	5	6	7	8
LXR-α	0	38	54	65	72	**77**	81	86
APO-B	0	39	54	64	73	**79**	83	86
CETP	0	45	69	**77**	82	86	90	93
PLG	0	35	47	56	64	72	**79**	84

Figure Legends

Fig. 1. Likelihood ratios under a fast- versus slow-mutation regime for genomic intervals containing apo-B exon 19 (Panel A), CETP exon 8 (Panel B), LXR-α exon 3 (Panel C), and plasminogen exon 6 (Panel D). The x-axis represents the position in the multiple alignment consensus sequence, the y-axis the log likelihood ratio at that position. The plot is smoothed by means of a 20%-trimmed mean over the 50-base window centered at each aligned site. A lower ratio indicates a higher degree of constraint on mutability of that site. The position of the exon in each sequence is shown by the purple line under the green ratio curve. The LXR-α plot contains a fragment of an additional exon at the right end of the plot. Panels a-d show the VISTA conservation plots for the corresponding orthologous regions in human and mouse. The y-axis represents the percentage of sequence conservation and the blue area the position of the exon.

Fig. 2. Analysis of the apo(a) 1.6 kb region. Panel A: likelihood ratios under a fast- versus a slow-mutation regime for the genomic interval containing the apo(a) exon 1 and 5'-flanking sequence. The x-axis represents the position in the multiple alignment consensus sequence, the y-axis the log likelihood ratio at that position. The plot is smoothed using a 20%-trimmed mean over the 50-base window centered at each aligned site. The position of the exon (E) is shown by the purple arrow. The blue (C1-C10) and red (N1-N7) rectangles show the sequence intervals that were selected as the representatives of conserved and non-conserved regions, respectively. Delineation of regions C9 and C10 was based on previous knowledge of their functional role. These

regions were investigated by gel-shift and transfection analysis. Panel B: electrophoretic mobility shift patterns of the conserved (C1-10) and non-conserved intervals (N1-7). Region C10, being too long (97 bp) for gel-shift analysis, was split into two 60-bp overlapping oligonucleotides (C10.1 and C10.2). Panel C: densitometric analysis of the electrophoretic mobility shift patterns. The y-axis represents the amount of shifted oligonucleotide normalized to the total amount of oligonucleotide in the lane. Columns and error bars represent the average and 1 standard deviation, respectively, of 5 (C1-C10.2) and 3 (N1-8) independent gel-shift experiments. The dotted line indicates the median electrophoretic mobility shift of the conserved and non-conserved regions.

Fig. 3. Transfection analysis of conserved and non-conserved regions in the 1.6 kb apo(a) region. The column bars represent expression values of luciferase reporter vectors carrying individual deletions from the whole 1.6 kb apo(a) region of one of the conserved (blue columns) or non-conserved (red columns) regions. Luciferase values are normalized by β-galactosidase expression and reported as percentages of the expression of the whole 1.6 kb apo(a) region. Each column represents the average of three independent transfection experiments, each determined in quadruplicate. Error bars indicate 1 standard deviation. The dotted line indicates the median expression of the conserved and non-conserved regions.